UMass Robust 2005:
Using Mixtures of Relevance Models for Query Expansion

Donald Metzler, Fernando Diaz, Trevor Strohman, W. B. Croft

Center for Intelligent Information Retrieval
University of Massachusetts, Amherst

Abstract

This paper describes the UMass TREC 2005 Robust Track experiments. We focus on approaches that use term proximity and pseudo-relevance feedback using external collections. Our results indicate both approaches are highly effective.

1 Introduction

For the 2005 Robust Track, we explore whether or not term proximity information and advanced pseudo-relevance feedback methods can be used to achieve good effectivness on a challenging query set.

All experiments used the Indri search engine [3], indexed the full AQUAINT collection of 1,033,461 documents, used a Porter Stemmer and a stopword list of 418 common terms. All runs are automatic.

2 Dependence Model

We use Metzler's dependence model formulation to exploit term proximity information, which been shown to significantly improve effectiveness over simple bag of words models [2]. The Indri query language can be used to express dependence model queries. This helps give an intuitive meaning to the model. For example, for topic 625, "arrests bombing wtc", the following Indri query ranks documents exactly as done by the dependence model:

```
#weight(0.8 #combine(arrests bombing wtc)
        0.1 #combine(#1(arrests bombing)
                     #1(bombing wtc)
                     #1(arrests bombing wtc))
        0.1 #combine(#uw8(arrests bombing)
                     #uw8(arrests wtc)
                     #uw8(bombing wtc)
                     #uw12(arrests bombing wtc)))
```

From this formulation we see that proximity information, in the form of exact phrases (`#1`) and unordered windows (`#uwN`) play a vital role in how documents are ranked.

3 Mixture of Relevance Models

Lavrenko's relevance models are a powerful way to construct a query model from a set of top ranked documents [1]. We generalize the idea to allow evidence to be incorporated from multiple collections. We take a Bayesian approach, and see that:

$$\begin{aligned} P(w|Q) &= \sum_{c \in \mathcal{C}} P(c|Q) P(w|Q,c) \\ &= \sum_{c \in \mathcal{C}} P(c|Q) \frac{\int_\theta P(w|\theta)P(Q|\theta)P(\theta|c)}{\sum_w \int_{\theta'} P(w|\theta')P(Q|\theta')P(\theta'|c)} \end{aligned}$$

In order to make evaluation of this expression more feasible, we follow Lavrenko [1] and approximate the integral by a summation over the models of the top ranked documents. We denote these models as $\mathcal{R}_c$,

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **2005**	2. REPORT TYPE	3. DATES COVERED **00-00-2005 to 00-00-2005**

4. TITLE AND SUBTITLE **UMass Robust 2005: Using Mixtures of Relevance Models for Query Expansion**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Center for Intelligent Information Retrieval,Department of Computer Science,University of Massachusetts Amherst,Amherst,MA,01003-9264**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**2**	

where the subscript indicates the collection. Furthermore, we also assume that $P(\theta|c) = \frac{1}{|\mathcal{R}_c|}$ and that $P(c|Q) = P(c)$ for all Q, which implies the mixture weights are equal for every query. Better distributional assumptions for $P(\theta|c)$ and actually computing $P(c|Q)$ may lead to better estimates, but is left as future work. Under these simplifying assumptions, we get the following estimate for our query model:

$$P(w|Q) = \sum_{c \in \mathcal{C}} P(c) \sum_{\theta \in \mathcal{R}_c} \frac{P(w|\theta)P(Q|\theta)}{\sum_{\theta' \in \mathcal{R}_c} P(Q|\theta')}$$

where we tune $|\mathcal{R}_c|$ and $P(c)$ on training data.

Now that we have a query model that combines evidence from multiple collections, we can use it for query expansion by adding the k most likely terms from the distribution $P(w|Q)$ to the original query.

In our experiments, we investigate mixing models from two collections, AQUAINT, and BIGNEWS, a collection of 6,160,058 TREC newswire articles we had on site.

4 Effectiveness Prediction

For predicting query effectiveness, we used a variant of the clarity measure, known as ranked list clarity [4]. Further details are omitted due to space constraints.

5 Results

The results of our official runs are given in Tables 1 and 2. Both the `indri05RdmT` and `indri05RdmD` runs are dependence model only runs. The `indri05RdmeT` and `indri05RdmeD` runs use a dependence model and mixture of relevance models with $P(bignews) = 1$, $P(aquaint) = 0$. Finally, the `indri05RdmmT` run uses the same formulation, except assumes $P(bignews) = 0.6$ and $P(aquaint) = 0.4$.

As we see, the dependence model results in a strong baseline and, when combined with mixture of relevance model expansion, produces very effective results for both title and description queries.

Run ID	MAP	GMAP	Area
indri05RdmT	0.2159	0.1354	1.4250
indri05RdmeT	0.3204	0.1967	2.3777
indri05RdmmT	0.3323	0.2061	2.6330

Table 1: Summary of Robust Track title only runs.

Run ID	MAP	GMAP	Area
indri05RdmD	0.1996	0.1015	0.9016
indri05RdmeD	0.2818	0.1611	1.9899

Table 2: Summary of Robust Track description only runs.

Acknowledgments

This work was supported in part by the CIIR, in part by NSF grant #CNS-0454018, in part by the DARPA, in part by ARDA and NSF grant #CCF-0205575, in part by NSF grant #IIS-0527159, and in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are the authors and do not necessarily reflect those of the sponsor.

References

[1] Victor Lavrenko and W. Bruce Croft. Relevance-based language models. In *Proceedings of SIGIR 2001*, pages 120–127, 2001.

[2] Donald Metzler and W. Bruce Croft. A markov random field model for term dependencies. In *Proceedings of SIGIR 2005*, pages 472–479, 2005.

[3] Trevor Strohman, Donald Metzler, Howard Turtle, and W. B. Croft. Indri: A language model-based serach engine for complex queries. In *Proceedings of the International Conference on Intelligence Analysis*, 2004.

[4] Steve Cronen Townsend, Yun Zhou, and W. Bruce Croft. Precision prediction based on ranked list coherence. *Information Retrieval (in submission)*, 2005.